リングネットワークにおける λコンピューティング環境に適した 共有メモリアーキテクチャの設計と評価

大阪大学大学院情報科学研究科
村田研究室　久保貴司

発表内容

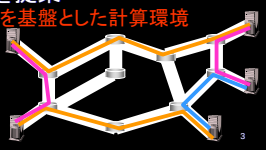
- 研究の背景と目的
  - λコンピューティング環境
  - 共有メモリアーキテクチャの設計と評価
- 設計
  - トポロジ
  - メモリアクセスモデル
  - キャッシュプロトコル
  - キャッシュとメモリの一貫性制御
- モデル化と解析
- 評価
- まとめ

背景

- グリッドコンピューティング技術への期待の高まり
- TCP/IPは分散並列計算には不向き
  - パケット処理や再送制御による遅延などのオーバヘッドが発生

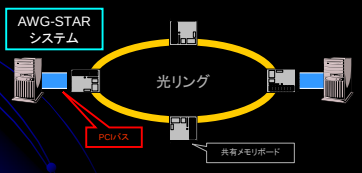
  ⇒大量のデータ交換を行う大規模計算への応用では十分な計算性能を達成することは困難
- λコンピューティング環境を提案
  - 高速・高信頼なネットワークを基盤とした計算環境

  ⇒ノード計算機間に波長パスを設定し，専用の通信路として用いる



従来の研究

- 関連研究[1–3]におけるλコンピューティング環境の実現
  - NTTフォトニクス研究所が開発した情報共有システム（AWG-STARシステム）を利用
- 並列計算アプリケーションの実行時間に基づいて評価
  - 分散計算を行うMPIやOpenMPIの設計と実装を行い，実際のアプリケーションを用いてその性能を評価
  - システムの各処理にどれだけ時間がかかっているのか不明



研究の目的

- λコンピューティング環境のシステムの定量的な評価
- λコンピューティング環境における計算性能
  - ネットワーク特性や共有メモリシステムにより影響
- λコンピューティング環境に適した共有メモリアーキテクチャの設計と評価
  - 設計
    - ハードウェア制約の考慮
    - 共有メモリアーキテクチャの特徴を決定する要因の検討
    - キャッシュー貫性制御に焦点を当てた動作の検討
  - 評価
    - ノード計算機のモデル化と解析
    - 計算スループットを指標として評価

共有メモリアーキテクチャの設計

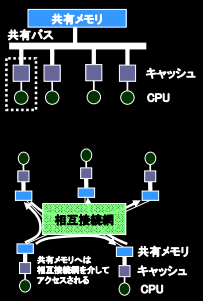
- A.トポロジ
  - リングトポロジ
  - メッシュトポロジ
- B.メモリアクセスモデル
  - Uniform Memory Access (UMA)
  - Non Uniform Memory Access (NUMA)
- C.キャッシュー貫性制御
  - キャッシュプロトコル
    - スヌープ法
    - ディレクトリ法
  - キャッシュとメモリ間の一貫性制御
    - 無効型と更新型
    - ライトスルーとライトバック

A.トポロジ

- トポロジ
  - リングトポロジ
    - 少数の波長で作成実現可能
    - ブロードキャスト時にデータの複製が不要
    - データ共有における伝播遅延・ノードの処理遅延は大
  - メッシュトポロジ
    - 多数の波長が必要
    - ブロードキャスト時にデータの複製が必要
    - データ共有における伝播遅延・ノードの処理遅延は小
- ハードウェア制約の考慮
  - WDM技術の発達により1つの物理回線に複数の波長を多重化することが可能
  - 多数の波長を利用するには多数のインターフェースが必要
    - 制御の複雑化
    - システムのコストが高

メモリアクセスモデル

- UMAモデル
  - 全てのCPUが同じ共有メモリを持つ
  - 共有メモリへ共有バスを介してアクセス
  - 任意のアドレスに対してCPUのメモリアクセス時間が同じ
  - 共有バスの競合
- NUMAモデル
  - 各CPUはキャッシュと共有メモリを持つ
  - それぞれの共有メモリは異なるアドレスを持つ
  - 他のCPUが持つ共有メモリへ相互接続網を介してアクセス
  - アドレスによってCPUのアクセス時間が異なる



キャッシュー貫性制御(1)

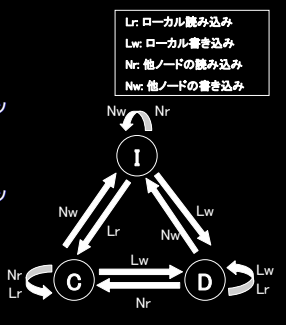
- キャッシュプロトコル
  - スヌープ法
    - 共有メディアを監視
      - システムの挙動を把握
  - ディレクトリ法
    - どのキャッシュがどのキャッシュラインを保持するかという情報を各ノードのディレクトリに記憶
- キャッシュとメモリ間の一貫性制御
  - キャッシュを一致させる方法
    - 書き込み時に該当キャッシュを無効化する無効化型
    - 書き込み時に該当キャッシュを更新する更新型
  - キャッシュを一致させるタイミング
    - 更新されたキャッシュに対して読み込みアクセスがあった場合に書き戻しを行うライトバック
    - キャッシュが更新されるたびに書き戻しを行うライトスルー

キャッシュー貫性制御(2)

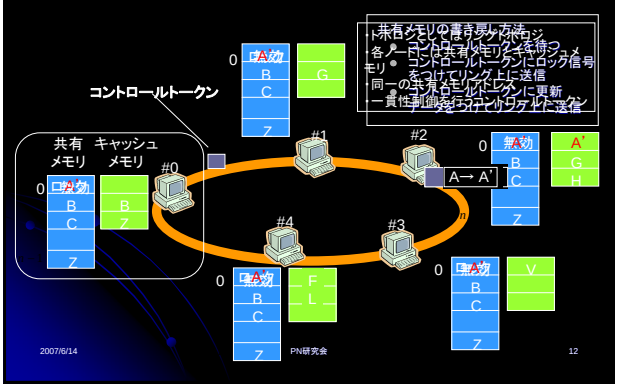
- 本研究では
  - キャッシュプロトコル：スヌープ法
    - リングトポロジを取るため，ネットワーク上のデータを監視することにより制御が簡単
    - トークンリングとし，トークンを監視することにより実現
  - キャッシュー貫性制御：無効化型ライトバック
    - 書き込みが行われたキャッシュを無効化するため，最新の情報を保持するキャッシュを記憶する必要がない
    - 書き込みが行われても，すぐに共有メモリへ書き戻しを行わず，読み込みが発生したときのみ共有メモリへ書き戻しを行う為，ネットワークの使用頻度が低い

キャッシュー貫性制御の内部状態

- キャッシュの状態
  - Clean状態
    - 共有メモリとCPUのキャッシュが一致
  - Dirty状態
    - 共有メモリとCPUのキャッシュが不一致
  - Invalid状態
    - CPUのキャッシュが無効



リングUMAアーキテクチャ



複数の波長パスの利用

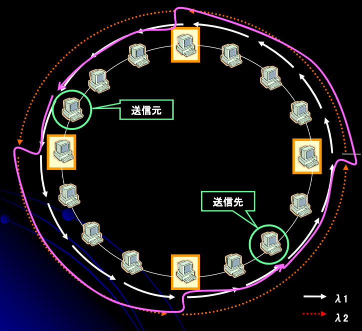
- リングトポロジでは，全ノードを経由するため処理遅延が大きくなる

↓

- 複数の波長を用いてノードをカットスルーし，処理遅延を小さく抑える

2007/6/14 PN研究会 13

波長パスの設定



- 波長経路が全てのノードを経由するように張る
- 2波長扱うことのできるノードを決定する
  - 分割時間数が均等になるように配置
- 波長1，及び波長2を扱うことのできるノードに張る
- 送信先アドレスによって決定される

2007/6/14 PN研究会 14

モデル化と解析

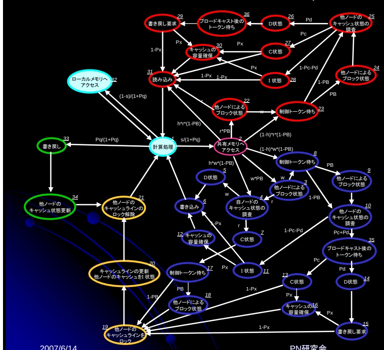
- 共有メモリアーキテクチャの動作に基づき状態遷移図を作成
  - ノード計算機のCPUの状態に着目
    - 計算処理を行っている状態
    - ローカルメモリにアクセスしている状態
    - キャッシュ一貫性制御を行っている状態　など
- 解析にはセミ･マルコフ過程を利用
  - 状態の滞在時間を任意に設定可能
  - 各状態の定常状態確率を導出

$$P_i = \frac{\eta_i \pi_i}{\sum_{j=1}^{K} \eta_j \pi_j}$$

P_i：状態 i の定常状態確率
η_i：状態 i の滞在時間
π_i：離散マルコフ連鎖時の状態 i の定常状態確率
K：状態数

2007/6/14 PN研究会 15

UMAモデル：2波長の状態遷移図



1	0.002	19	τ
2	0	20	τ
3	3τ / 2 + T/2	21	T
4	t1	22	3τ / 2 + T/2
5	0	23	0.5τ
6	t1	24	3τ / 2 + T/2
7	0	25	T
8	0.5τ	26	0
9	3τ / 2 + T/2	27	0
10	T	28	0
11	0	29	2τ
12	t2	30	t2
13	0	31	t1
14	0	32	h*t1+(1-h)(t1+t2)
15	2τ	33	t1+t2
16	t2	34	2τ
17	0.5τ	35	τ−T
18	3τ / 2 + T/2	36	τ−T

共有メモリへのリードアクセス時に通過する状態
共有メモリへのライトアクセス時に通過する状態
キャッシュ一貫性制御を行っている状態
共有メモリへ書き戻しを行っている状態
ローカルで処理を行っている状態

2007/6/14 PN研究会 16

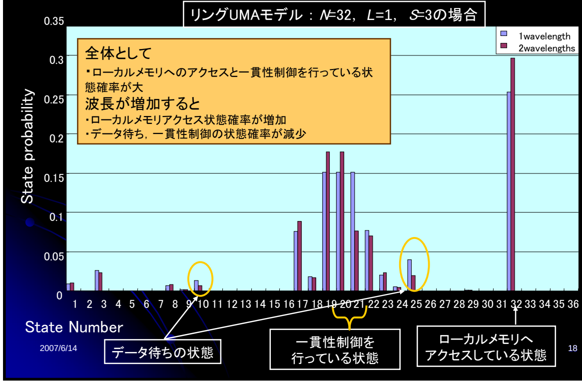
評価のための数値例

- 計算機ノードのスペック
  - CPU－キャッシュ間の遅延 10ns
  - キャッシュ－メモリ間の遅延 1us
  - キャッシュ容量　1024 KB
  - キャッシュラインサイズ 4 KB
- ネットワーク
  - インターフェイスでの処理遅延 3us
  - 帯域 10Gbps
  - 伝播遅延 5 us/km
- その他，必要となるパラメータ
  - キャッシュヒット率 0.95
  - CPUが発行する命令頻度
    - LAD命令 15%，STORE命令 5%，その他（レジストリ間計算） 80%
- 解析のパラメータ
  - ノード数 *N*(4 ～ 64)
  - 光リングの長さ *L* km (0.01km ～ 100km)
  - 共有メモリへのアクセス頻度 *S*(10^(-1) ～ 10^(-6))

2007/6/14 PN研究会 17

状態確率の分布

リングUMAモデル：*N*=32，*L*=1，*S*=3の場合



2007/6/14 PN研究会 18

スライド 16

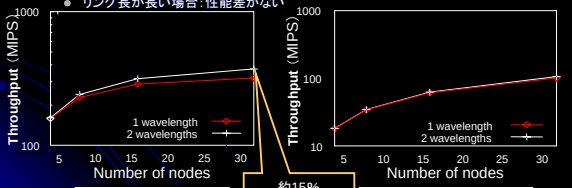
t3

UMA・NUMAで
2波長の場合の状態遷移図である.
1波長の場合の状態遷移図である.

t-kubo, 2007/06/03

計算スループットによる評価

- 全体として
  - リング長が短く，共有メモリへのアクセス頻度が低い場合，高計算スループット
  - リング長が長い，又は共有メモリへのアクセス頻度が高い場合，低計算スループット
- 波長数による性能差
  - リング長が短い場合：ノード数が増加すると性能が向上
  - リング長が長い場合：性能差がない



リングUMAモデル
L=1km, N=32, S=3の場合

約15%
性能が向上

リングUMAモデル
L=100km, N=32, S=3の場合

まとめと今後の課題

- まとめ
  - 共有メモリアーキテクチャの設計
    - リングUMAアーキテクチャ（1波長，及び2波長）
    - リングNUMAアーキテクチャ（1波長，及び2波長）
  - 計算スループットを指標とした共有メモリアーキテクチャの評価
    - UMA・NUMA共にほぼ同じ性能
    - リング長が短く共有メモリへのアクセス頻度が低い場合，高計算スループット
    - リング長が長い，又は共有メモリへのアクセス頻度が高い場合，低計算スループット
  - 波長数による性能の違い
    - UMA・NUMA共に同じ傾向
    - リング長が短い場合
      - 共有メモリへのアクセス頻度が高いと性能差が生じる
      - ノード数が多いほど複数波長を用いたアーキテクチャの方が性能はよくなる
    - リング長が長い場合
      - 共有メモリへのアクセス頻度，ノード数によらず性能の差がない
      - ノードの処理遅延に比べ，伝播遅延が大きいためと考えられる
- 今後の課題
  - メッシュトポロジにおける共有メモリアーキテクチャの性能評価
    - 伝播遅延が小さいため，高計算スループットが期待